SGOP Project Series

The SGOP Project

Firenze August 2019
C.Baffa, E.Giani, E.Pancino

Arcetri Technical Report 8/2018

Abstract

The large amount of astrometric data available with the publication of Gaia catalogs 1 and 2 permits statistical analysis previously impossible. We seek a new way to explore this wealth by the use of the empirical concept the Star Group with Observational Peculiarity (SGOP). The concept of SGOP, in our opinion, can be a useful tool to explore the path connecting the large Gaia database to a catalog of star association candidates, selected on objective observational properties basis .

1 Introduction

With the Gaia database readily available with VO tools as Aladin[1] or Topcat[2], it is easy to find some evidence of the presence of conspicuous star clusters. In Figure 1 we can easily see the signature on Gaia DR2 ([4], [5]) data of M6 cluster presence.

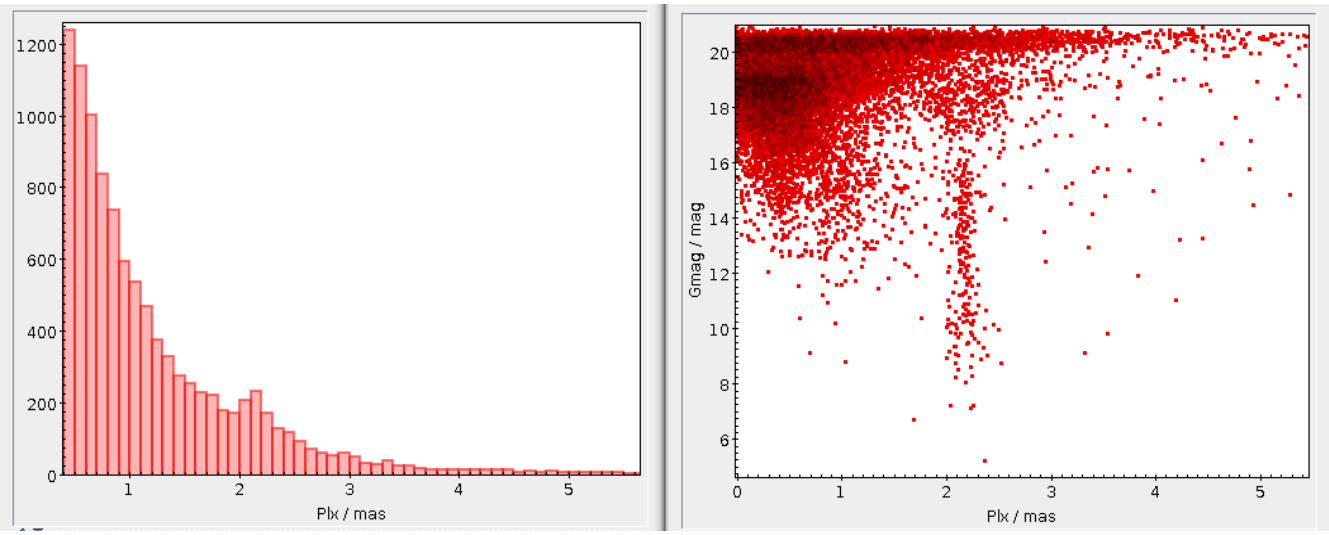

Figure 1: Left panel Gaia DR2 Parallaxes Histogram around M6 center. Right panel plot of parallaxes versus GMag for the same field

If we compare the same plots taken at 20' distance from cluster center along galactic latitude, the effect is still more remarkable. We note 20' is the core size of M6.

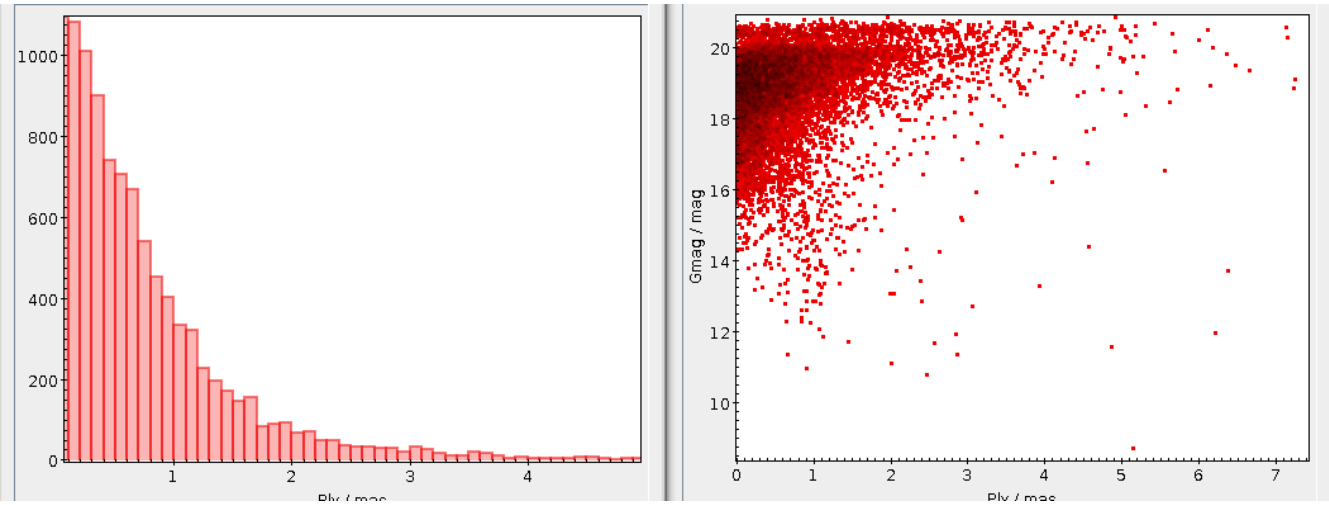

Figure 2: Left panel Gaia DR2 Parallaxes Histogram 20' from M6. Right panel plot of parallaxes versus GMag for the same field

1 This work has made use of "Aladin sky atlas" developed at CDS, Strasbourg Observatory, France [3]

2 Topcat [1] and Stilts [2] can be obtained from Starlink web site: http://www.starlink.ac.uk/topcat/ and http://www.starlink.ac.uk/stilts/

If we select the peculiar star group from right panel of Figure 1 and draw again the same plots using different colors for the different sets, the peculiarities are still more evident. We select all stars with 1.8mas<plx< 2.4mas (box set, blue in the plot), and from this set the “spear like” group (M6 set, green). In Figure 3, the central panel shows the Parallax histogram of All stars (gray), Parallax selected (blue) and spear like group (green). Left panel shows the two latter groups in M_G/B-G color plane, right panel shows them in Ra/Dec proper motion plane.

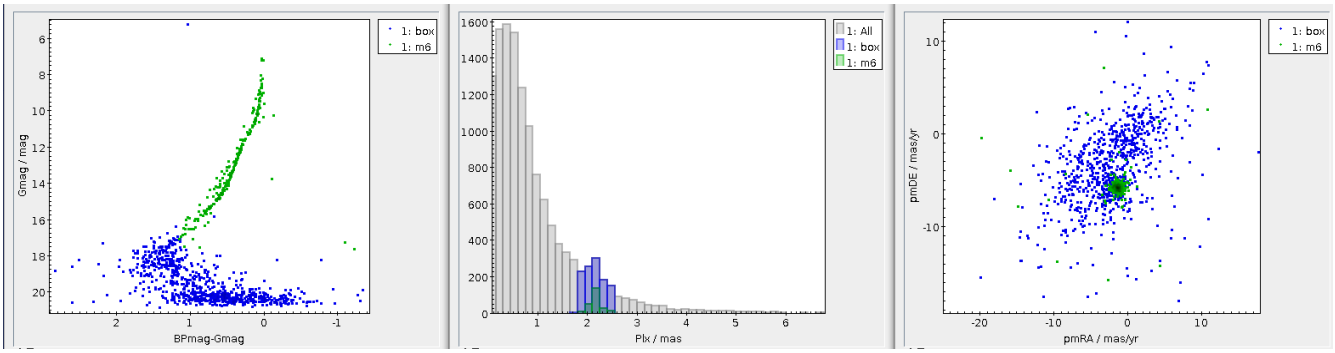

Figure 3: Refinement of SGOP selection. Central panel shows the parallax histogram of all stars (gray), parallax selected group (blue) and spear like group (green). Left panel shows the two latter groups in M_G/B-G color plane, right panel shows them in Ra/Dec proper motion plane. Here the purple group is represented in blue

From the above evidences we devised a potentially useful tool. We defined an empirical concept: the Star Groups with Observational Peculiarities (SGOP). In our concept a SGOP is a group of stars which collectively shows one or more peculiar characteristics statistically different from the local environment stars groups (LESG). In the following we will propose some different types of SGOP anomalies.

The SGOP approach is complementary to other current Gaia DR2 base cluster search as that devised by Kounkel and Covey ([8]). That work has an holistic approach on 5 kinetic coordinates, which are available with enough quality only for a small sub-sample of total Gaia catalog. That approach gives a wealth of new information, but cannot reach farther cluster nor can collect information on fainter star. Our approach cover such areas and moreover having a more direct statistical approach ,in principle can also give clues on completeness and introduced biases.

2 SGOP types

From an *a priori* approach we can define a preliminary list of peculiarities to be used to start the study (in 1). Of course more types can be defined.

It can easily seen that for many fields where SGOPs are fainter, such peculiarities can be difficult to spot. So we have devised some procedures to assess the statistical significance of an anomaly.

Some peculiarities (for instance 2,3 and 4) can be used to obtain a centroid of the SGOP, so the detection procedure can be iterated, in order to obtain a better S/N.

SGOP type	Observational Anomaly	Notes
1 - Density	Increased density from Plx histogram	Easy
2 - Proper Motion	A peak distribution in Ra/Dec Proper Motion	Easy (bi-dimensional)
3 - Color - Color	A peculiar distribution in color/color plane	Easy (bi-dimensional)
4 - Luminosity	A peculiar distribution in Mag histogram	Easy
5 - Color-Magnitude	A peculiar distribution in color/Mag plane	Easy (bi-dimensional)
6 - Motion azimuth	A peak distribution in Proper Motion azimuth	Easy
7 - Parallax-Magnitude	A peculiar distribution in Plx/mag plane	Less informative (noisy plx)
8 - Radial velocity	A peculiar distribution in Plx/RV plane	Difficult (scarcity of data)

Table 1: Preliminary list of SGOP types.

3 Searching SGOPs

We have defined the SGOPs as statistical anomalies over the general background of Milk Way. Due to the huge original sample (Gaia Dr2) we need to carefully plan our search methodology to maximize the probability to find them. Later on we will give initial values for all quantities we need for this analysis.

Our analysis will be performed on Gaia 2 catalog and the star sample will be selected in the space defined by galactic coordinates, l and b, and parallax. We chose the galactic coordinates as the natural reference frame for galactic objects as clusters. A glance to open cluster spatial distribution (Figure 7) makes this point evident.

Considering the Gaia parallax values, we found we have at least three different search areas, with different search needs. The three areas are defined along the distance axis. We will discuss later the boundaries locations. The areas we consider are:

1. *The nearest area (or solar neighborhood)*. This area cannot be easily examined with the spherical shell model we discuss below. We need to revert to a Cartesian approach.
2. *The intermediate area (at intermediate distances)*. At first we will concentrate on this regime.
3. *The farther area (farther distances)*, where the measured parallaxes and relative error are of the same magnitude.

We plan to analyze, at first, area 2 and later area 3. Here we will discuss area #2, the intermediate area.

For the SGOP quest, our idea is to examine a spherical shell (or a portion of it) at time. The shell width should be comparable with the size of star groups we already know, convoluted with the parallax typical error. For the SGOP quest we plan to divide the shell in small cuboids, each with linear dimensions comparable with the shell width.

The edges of the cuboids are along the axis of the the l, b and Parallax space. It is easy now to analyze a cuboid at time. The setting can be visualized in Figure 4. The use of cuboid term comes from the fact that the faces perpendicular to Parallax axis are not plane. The use of cuboids instead of the much easier VO cone search comes from the attempt not to dilute the SGOP peculiar quantities by merging different sky areas.

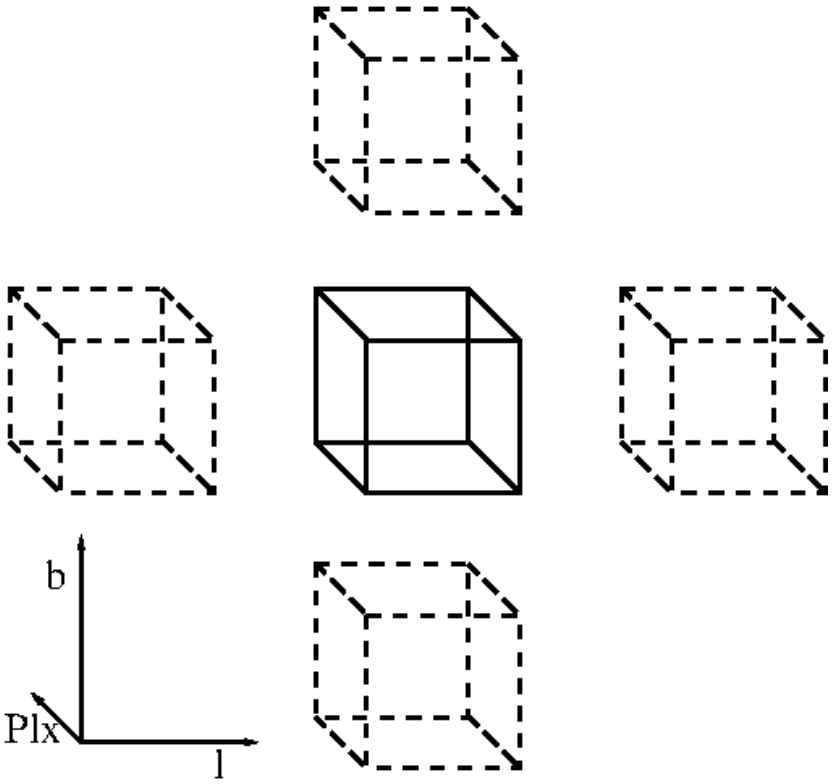

Figure 4: SGOP quest setting. Each cuboid is cut from a spherical shell in the l,b and Plx space. The three dimensions of each cuboid are of the same order of magnitude of already known star cluster. The star sample under analysis is inside the central cube, the dotted cubes acts as reference.

To obtain a better detection probability, we define empty field values for the quantities we are analyzing. For instance, for sgop1, the characterizing quantity is the histogram of sources counts along the parallax axis. We start computing the histogram of the cuboid under analysis. We compute *median* of the histograms of four references fields, to define a *field histogram*, to be compared with the *cluster histogram*. If the difference is statistically significant we can mark the cuboid under analysis as sgop1. This procedure can be effective also when the cluster/field density contrast is very low. In Figure 4 we can see the cuboid under analysis (solid line) and the four reference field (dashed lines). To minimize

the effect of larger star group and/or mismatch between star group centroid and the center of cuboids, we choose references not exactly adjacent to the central cuboids.

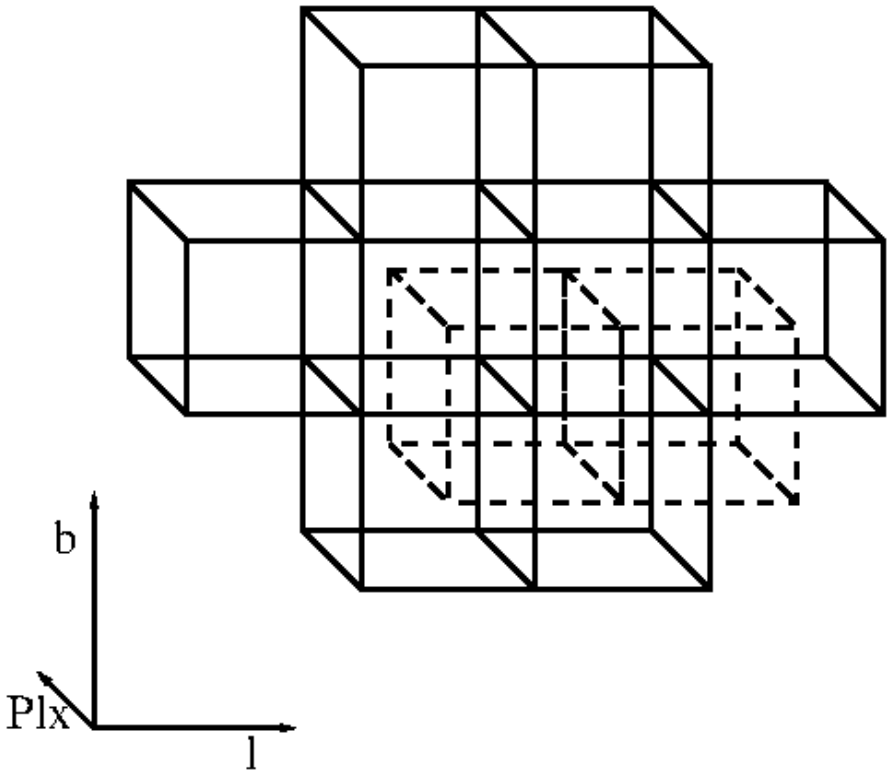

Figure 5: To increase the SGOPs detection probability, each spherical shell is scanned twice, with the second cuboids shifted in b and l by half step (dashed cuboids).

To increase the SGOPs detection probability, each spherical shell is scanned twice, with the second cuboids shifted in b and l by half step (see Figure 5). Also each spherical shell increases in radius by half their width. This approach multiply the number of operations by a factor of 4, but greatly increase the chance to find a SGOP.

There is the possibility to detect multiple times the same SGOP, so we will devise an appropriate *coalescence* procedure.

4 SGOP search starting values

The full Gaia catalog is a very large field for the SGOP search described above. We have already found there are at least three regimes (see list in chapter 3) and we need to define many parameters in order to start a meaningful search. Most of these parameters are related to the final scope of SGOPs search: we aim to find candidates for galactic associations. As a consequence we need the order of magnitude of their physical parameters. As a guide we used an extensive catalog easily available in VO format ([6], hereafter Kharchenko).

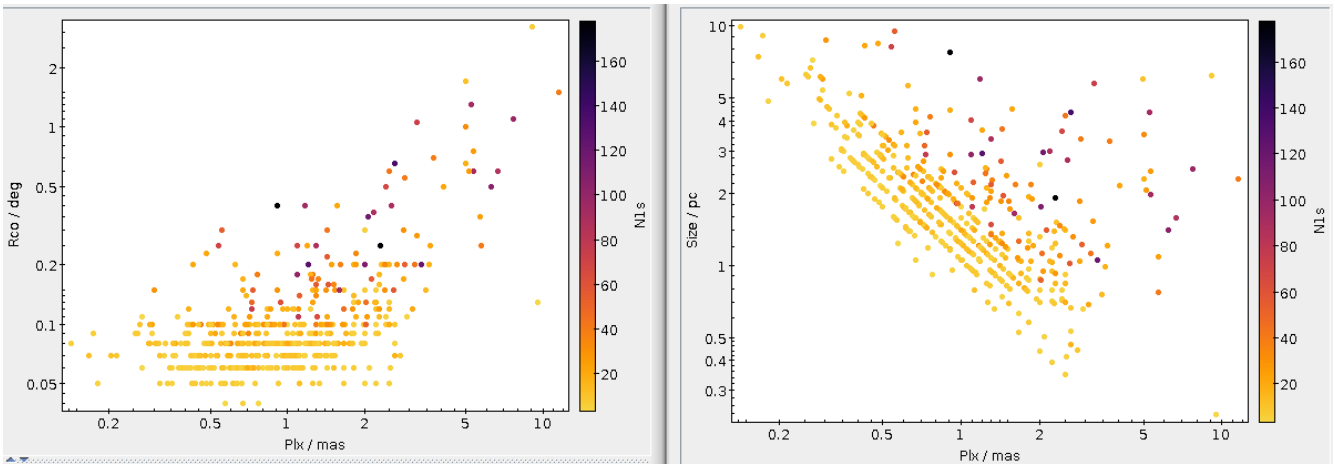

Figure 6: Distribution of size versus computes parallaxes for Karchenko catalog of galactic open clusters. Left panel apparent angular size versus Plx, right panel physical size versus Plx

4.1 cuboids size

In Figure 6 we report plots of Kharchenko's open clusters' linear dimensions as a function of reported parallaxes. For linear dimension we used the catalog Rco (core Radius), in parsec, and we computed the expected parallaxes from the catalog distance field. We used a quick-and-dirty approach as we need only order of magnitude estimates, therefore we do not perform a critical analysis on Kharchenko catalog reported size.

From Figure 6 we can spot there are some parameter space for cluster detection using Gaia data (for instance the lower left corner of right panel). From the right panel plot we can assume a typical cluster core radius between 1pc and 2pc (evaluated at 1000pc). This translates to an angular size of about 0.12º (8'). We can scale linearly this value towards larger parallaxes, (to 24' at 300pc). To assume the same behavior towards larger distances, where the magnitude limit of Gaia catalog is more important, is probably more troublesome, and needs a check on real data. We need to consider also the further shells require most work, so a careful plan is strongly needed.

To reduce S/N dilution we should use also, on radial direction, the smallest useful size. For example for the shell at plx=3mas this would translate to a very thin plx interval (0.01mas). However the relatively large Gaia DR2 measurement error (we assume 0.3mas) force us to use the theoretical value convoluted to the measurement error. In practice we use Δplx=0.33mas: a very elongated cuboids!

4.2 cuboids spatial distribution

Figure 7 shows the spatial distribution of Kharchenko's open clusters, in galactic coordinates. It is evident there are only very few at latitude larger than 30º. Therefore, for the first iteration we plan to analyze only the -30º<b<30º zone.

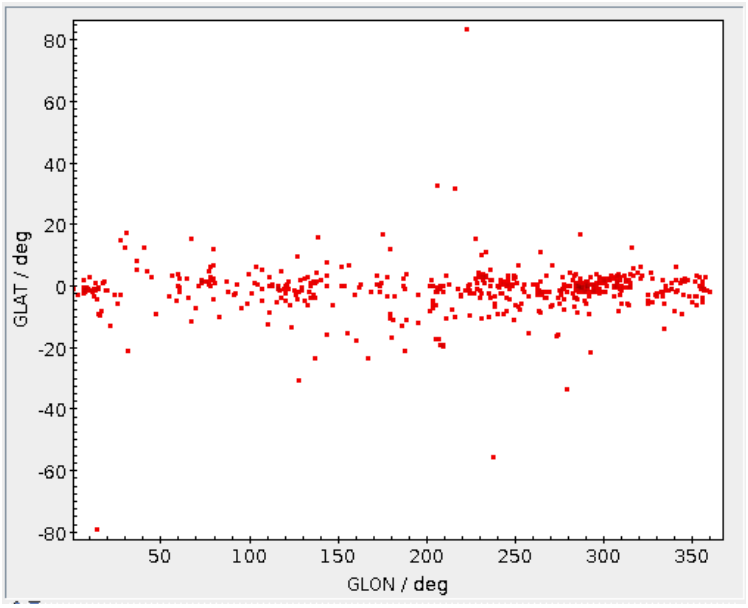

Figure 7: Spatial distribution, in galactic coordinates of Kharchenko catalog cluster. It is evident the strong increase in density towards the Galactic plane

For the circumpolar regions we expect to use a different partitioning, both in spatial and in parallax. A different spatial partitioning is needed to maintain a more or less constant size of search areas, while a different parallax partition schema (and maybe none at all) is intended to save computing time in a very cluster-poor region (the halo region).

4.3 cuboid numbers

With the search parameter defined in the previous sections we can give some estimates of the required work. The cuboids size and number can be seen in 2. As expected, the number of cuboid to inspect rises sharply with distance: the last shell accounts for 2/3 of the total!

We need also to consider that, using the previously described procedure, the real number of cuboid to consider is 4 times the reported one.

The total number of main cuboids listed in 2 is about 7.5e6. This number, while large is absolutely not outside a medium size research effort, if appropriate software tools are used.

Parallax (mas)	Size in l,b (primes)	Size in Plx (mas)	Number of cuboids
3.00	36.00	0.33	60000
2.66	32.00	0.33	76000
2.33	28.00	0.33	99000
2.00	24.00	0.33	135000
1.66	20.00	0.33	194000
1.33	16.00	0.33	304000
1.00	12.00	0.33	540000
0.66	8.00	0.33	1215000
0.33	4.00	0.33	4860000

Table 2: Size and number of first iteration of SGOP search plan.

We used the Stilts java program to handle the large Gaia2 database operations, while we plan to write some custom code to perform more specific analysis. The use of optimized code to cope with such large database does not avoid long execution time: on a high level desktop the steps 2-4 of Chapter 5 took from many days to few weeks each.

4.4 SGOPs detection

At the end of extraction process data (depending of SGOP variety) a 1D or a 2D histogram will be confronted with the median of histograms of reference zones.

Our case is a null hypothesis problem and can be handled with some classical approach or a Bayesian one. For mono-dimensional SGOP type a classical χ^2 test is probably the best choice. With bidimensional types, we plan to use the procedure for histograms comparison detailed in [7], based on a classical approach.

Our case does not present scarcity of data, so we do not foresee the need for the heavier Bayesian approach, but real data can prove otherwise.

4.5 SGOPs coalescence

The multiplicity of same sky analysis devised in Chapter 3 requires a procedure to avoid multiple entries for the same SGOP. From the detail of previous analysis, a preliminary very simple algorithm can be devised.

We plan to coalesce SGOP whose centroid have a distance inferior to the size of the cuboids used for their detection. Some care will be taken for binary cluster, which this approach will identify as a single elongated cluster.

This euclidean distance approach seem now a simple and efficient approach. We will discover, as data will accumulate, if it will be able to cope with the variety of situation we will face.

5 Operations Plan

The first phases of SGOP search program can be described as follows:

1. Collection of Gaia DR2 csv files (>64K files). DONE
2. Conversion of Gaia DR2 csv files into fits files (>64K files). DONE
3. Merge of Gaia DR2 fits files into a single fits table with RV. DONE
4. Split of the single fits table into 360 one degree longitude, -30:+30 latitude files. DONE
5. Selections of main “avoidance zones”.
6. Development of single-zone python extraction procedure. DONE
7. Development of multiple-zone python extraction procedure. DONE
8. Development of single zone python SGOP values extraction. DONE
9. Development of single SGOP statistical detection program.
10. Development of multiple SGOP type statistical detection program.
11. Selection of a single sky area to test the method.
12. Application of SGOP search to a single shell for most of sky.
13. Application of SGOP search to multiple shells, up to penultimate one.
14. Application of SGOP to farthest shell.

Most of the code is written as bash procedure or python scripts.

All the code is collected in a git archive, accessible at

`https://gitlab.com/carlobaffa/gcluster`

We choose to apply a GPL2 license to our code.

6 Conclusions and Acknowledgments

The SGOP research plan presented above appear to be a promising course of action and comlementary to some current approaches to Gaia DR2 based clusters searches. It is evident it will require a substantial amount of work, in terms of coding, of computing power and of post processing refinement. However we foresee many benefits from the compilation of even a partial SGOP catalog and we commit ourselves to pursuit this effort as far as possible/productive.

One of us (CB) would like to acknowledge the fruitful discussions and suggestions of S.Casertano.

CC and EG thanks M. B. Taylor for his clarifications and suggestions on Topcat/Stilt use.

7 References

[1] Taylor, M. B., “TOPCAT & STIL: Starlink Table/VOTable Processing Software”, in ADASS 2004, ASP Conference Series, Vol. 347, 2005, p.29

[2] Taylor, M. B., “STILTS - A Package for Command-Line Processing of Tabular Data”, in ADASS 2005, ASP Conference Series, Vol. 351, 2006, p.666

[3] Bonnarel, F., Fernique, P., et al, “The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources”, A&AS, v.143, p.33-40, 2000

[4] Gaia Collaboration, et al., “The Gaia mission ”, A&A **595**, pp. A1.

[5] Gaia Collaboration, et al., “Gaia Data Release 2. Summary of the contents and survey properties.”, ArXiv e-prints 1804.09365.

[6] Kharchenko, N. V., et al, “Astrophysical parameters of Galactic open clusters”, 2005, A&A, **438**, pp.1163-1173

[7] Bityukov, S., Krasnikov, N., Nikitenko, A., Smirnova, V., 2013, "A method for statistical comparison of histograms", arXiv:1302.2651 [physics.data-an].

[8] Kounkel, M and Covey, K, 2019, “Untangling the Galaxy. I. Local Structure and Star Formation History of the Milky Way}”, AJ, **158**, 3, p.122

8 Index

Table of Contents

Abstract....2
1 Introduction....3
2 SGOP types....5
3 Searching SGOPs....5
4 SGOP search starting values....8
4.1 cuboids size....8
4.2 cuboids spatial distribution....9
4.3 cuboid numbers....9
4.4 SGOPs detection....10
4.5 SGOPs coalescence....10
5 Operations Plan....11
6 Conclusions and Acknowledgments....12
7 References....13
8 Index....14